基于多特征融合的金融领域科研合作推荐研究*

余传明[1] 龚雨田[1] 赵晓莉[1] 安 璐[2]
[1](中南财经政法大学信息与安全工程学院 武汉 430073)
[2](武汉大学信息管理学院 武汉 430072)

摘要:【目的】科研合作关系是一种重要的社会网络。为了促进科研合作,提高科研生产率,对金融领域的科研合作推荐模型进行研究。【方法】建立金融领域个人、机构和区域三个层面的科研合作网络,提出一种新的融合基于邻居节点和基于路径的网络特征的科研合作推荐模型,并从个人、机构和区域三个层面进行实证检验。【结果】通过对 2000 年到 2014 年刊载的 68 905 篇金融领域的文章进行分析并构建科研合作网络,在个人、机构和区域三个层面上,基于特征融合的链接预测方法的 AUC 值分别为 84.25%、87.34%和 91.84%,均高于基于邻居节点的算法和基于路径的算法的 AUC 值。【局限】在进行训练集和测试集选取的时候只按时间进行切分,有待使用更多的切分方式对实验结果进行优化。【结论】本文有助于金融科研领域的个人、机构和区域寻求合作对象,为进行科研网络的研究以及科研合作推荐的学者提供新的研究方法和思路。

关键词: 链接预测 科研合作推荐 科研合作网络 多特征融合
分类号: G350

科研合作网络是一种重要的社会网络。在科研合作网络中,通常将科研合作对象(包括个人、机构和区域等)抽象成为一个节点,将科研对象之间的合作关系抽象成为一条边。链接预测(Link Prediction)是社会网络分析的一个重要问题,其任务是根据已知的链接和节点的属性,来预测尚未连接的两个节点之间存在或产生链接的可能性[1-3]。链接预测既包括预测将来会产生的新链接(Future Links),也包括预测已经存在但尚未发现的链接(Exist but Unknown Links)[4]。基于链接预测技术进行科研合作推荐在国内外得到越来越多的关注,所采用的方法包括基于邻居节点的链接预测和基于路径的链路预测等。目前的研究存在以下问题:

(1) 往往只考虑基于邻居节点或者基于路径的网络特征,很少有将两者融合在一起,而这两类算法在进行链接预测的时候往往各有优势;

(2) 国内的研究大多局限于个人层面,在机构和区域层面少有涉及。

鉴于此,本文提出一种融合基于邻居节点和基于路径的网络特征的科研合作推荐模型,并从个人、机构和区域三个层面进行实证检验,以期为相关研究提供借鉴。

1 文献回顾

在社会网络分析领域,链接预测算法通常分为三种类型,即相似度算法、最大似然模型和概率方法[5]。相似度算法通过计算节点所共有的特征来定义相似度。节点相似度越高,在推荐过程中拥有越高的优先权。最大似然模型通过最大限度地提高网络结构的可

通讯作者:安璐,ORCID:0000-0002-5408-7135,E-mail:anlu2009@whu.edu.cn。
*本文系国家自然科学基金面上项目"大数据环境下基于领域知识获取与对齐的观点检索研究"(项目编号:71373286)、国家自然科学基金青年项目"突发公共卫生事件社交媒体信息主题演化与影响力建模"(项目编号:71603189)和教育部人文社会科学研究青年基金项目"突发公共卫生事件情境下社交媒体信息影响力模型与预测研究"(项目编号:16YJC870001)的研究成果之一。

能性，根据获取的规则和参数来计算所有推荐链路的可能性。概率方法则是用一组参数组合的概率模型估计待推荐的链接概率。由于目前最大似然模型和概率方法在处理大型网络时效率仍然较低[3, 6]，且本文研究的科研合作网络涉及到节点数量较多，因此采用相似度算法作为本文研究基础。

基于相似性的链接预测算法可以分为两类，即基于邻居节点的算法和基于路径的算法。基于邻居节点的链路预测方法包括共同邻居(Common Neighbors, CN)算法[7]、Admic/Adar(AA)算法[8]、Jaccard Coefficient (Jaccard)算法[9]和 Preferential Attachment(PA)算法[10]等。Chen 等较早地提出共同邻居算法，将共同邻居定义为两个节点间所共同拥有的连接节点数量[7]。节点之间拥有越多的共同邻居，则越有可能建立新的链接。AA 算法和 Jaccard 算法对 CN 算法进行了补充，对于稀有特征(Rare Features)赋予更大的权重。其基本假设是，两个拥有共同稀有兴趣的人可能更易成为朋友，且其成为朋友的可能性往往正比于所共有兴趣的稀有性。与 AA 算法不同，PA 算法则认为度数较高(即拥有更多邻居)的节点具有更高的可能性来建立新的链接，并且这种可能性与其度数的乘积成正比。基于路径的算法则包括 Shortest Path(SP)算法、Katz 算法[11]、FriendLink 算法[12]、Random Walk with Restart(RWR)算法[13]等。其中，SP 算法计算节点对之间的最短路径，并认为具有较短路径的节点对之间更容易建立链接。与 SP 算法仅仅考虑一条路径不同，其他算法则将多个路径综合加以考虑。例如，Katz 算法将节点之间所有的路径进行累计求和；FriendLink 算法则对不同长短的路径赋予不同的权重，认为那些具有独特路径(Unique Path)的节点之间更容易建立链接。

Yan 等较早地将链接预测方法应用到科研合作推荐之中，将 CN 算法、AA 算法、Jaccard 算法、PA 算法和 Katz 算法等应用到图书馆学情报学领域的科研合作推荐之中[14]。张斌将上述多种方法应用到包括文学、历史学、法学和教育学等在内的多个学科的科研合作者推荐之中[15]。刘萍等利用 LDA 主题模型进行科研合作推荐[16]。吕伟民等尝试将链接预测与机器学习结合，以提高科研合作推荐的精确度[17]。上述方法在多个领域的科研合作推荐之中取得了一定的成果，但多数研究是将基于邻居节点的算法和基于路径的算法孤立开来进行科研合作推荐。

2 研究方法

将4种基于邻居节点的预测方法(CN算法、Jaccard算法、AA 算法和 PA 算法)和 4 种基于路径的预测方法(SP 及改进最短路径算法、Katz 算法、RWR 算法和 FriendLink 算法)作为基线方法，与本文提出的融合方法进行比较。

2.1 基于改进最短路径的预测方法

最短路径算法作为几大经典算法之一，在计算机科学、运筹学等学科中一直是一个研究热点。之前的学者对该算法的研究解决了优化网络特征运行结构等一系列网络特征问题，推动最短路径算法越来越成熟。但是在对科研者进行合作推荐时需要考虑多个科研者的路径相似度问题，传统的最短路径算法并不能很好地解决这一问题，因此笔者考虑对最短路径算法进行改进，将科研者最短路径的相似性考虑进来，改进后的算法如公式(1)所示。

$$Path_Sim(i,j)=\left[\sum(A,B)\in shortestpath_{ij}\frac{1}{|P_A\cap P_B|}\right]-1 \quad (1)$$

其中，$Path_Sim(i,j)$表示节点i和节点j之间的最短路径相似度，(A,B)表示合作节点对，该节点对是节点i和节点j的最短路径上的节点，假设节点i到j的一条路径为$\{V_0=V_i,V_1,\cdots,V_{l-1},V_l=V_j\}$，则合作者对集合为$\{(V_i,V_j),(V_1,V_2),\cdots,(V_{l-2},V_{l-1}),(V_{l-1},V_l),(V_l,V_j)\}$。$P_A$表示节点$A$的论文集合，$|P_A\cap P_B|$表示节点$A$和节点$B$合著论文的数目。如果两个节点之间合作的次数越多，则该公式得出的结果值越大。

2.2 基于多特征融合的预测方法

本文将多种特征得出的相似度结果进行融合，形成一致性的数据模型。由于每种特征的计算结果在数量级上可能有很大差别，直接进行融合会导致数量级较大的特征占主导地位，从而造成不准确的结果，因此在进行特征融合之前，首先将所有的相似度结果进行 min-max 归一化，如公式(2)所示。

$$score_{norm}=\frac{score-\min(score)}{\max(score)-\min(score)} \quad (2)$$

本文采用线性组合方法，针对各种特征的相似度

构造融合模型(Proposed Hybrid Method, PHM)，如公式(3)所示。

$$score_{fused} = \alpha \times \max\{rel_i(score_{norm})\} + (1-\alpha) \times \max\{rel_j(score_{norm})\} \quad (3)$$

其中，$score_{fused}$ 是最终的相似度，作为推荐合作的依据；$rel_i(score_{norm})$ 表示基于路径的算法的计算结果，$\max\{rel_i(score_{norm})\}$ 表示基于路径算法中(SP及改进算法、Katz[11]算法、RWR[13]算法和 FriendLink[12]算法)最优的计算结果；$rel_j(score_{norm})$ 表示基于邻居节点的算法中(CN 算法、Jaccard 算法、AA 算法和 PA 算法)的计算结果，$\max\{rel_j(score_{norm})\}$ 表示基于邻居节点算法中最优的计算结果；α 值为基于路径算法的权重，其值在 0.1 和 0.9 之间采取步长为 0.1 进行动态调整。

3　实验过程与结果分析

3.1　数据获取与预处理

选取中国知网期刊数据库金融类目 CSSCI 选项下所包含的 68 905 篇论文(2000 年–2014 年)作为数据源，从个人、机构以及区域三个层次构建科研合作网络。假设有一篇论文发表于 2014 年，在该论文中有三个合著者 author1、author2、author3，则构建三个科研作者合作对，即<author1,author2,2014>、<author1,author3,2014>和<author2,author3,2014>。同理，如果作者 1 和作者 2 属于机构 institution1，作者 3 属于机构 institution2，则机构 1 和机构 2 之间形成一条科研机构合作对，形成机构对<institution1,institution2,2014>。区域层次的科研合作网络构建与此类似。

将数据集分成两部分，即训练集和测试集。其中，训练集是 2000 年–2013 年的数据，测试集是 2014 年的数据。在区域合作层次，某些国外地区以及因为年代和信息不完全问题导致的无法判断所属地区的数据在实验中被剔除。所使用三个层面的数据集如表 1 所示。

3.2　基于邻居节点和路径的预测方法

将科研对象(个人、机构、区域) A 和 B 抽象为节点(下文同)，两者拥有共同的合作对象 C，则 A、B 有可能建立新的合作。A、B 拥有的共同合作对象越多，则其建立新的合作关系的可能性越高。本文所使用的基于邻居节点的预测方法包括 4 种: CN 算法[7]; Jaccard 算法[8]; AA 算法[9]; PA 算法[10]。

表 1　实验数据集

	个人层面	机构层面	区域层面
总集合	作者数: 4 123 合作链接数: 7 096	机构数: 3 383 合作链接数: 12 336	区域数: 46 合作链接数: 411
训练集	作者数: 4 049 合作链接数: 6 119	机构数: 3 289 合作链接数: 11 241	区域数: 46 合作链接数: 400
测试集	作者数: 1 080 合作链接数: 864	机构数: 1 109 合作链接数: 1 553	区域数: 36 合作链接数: 171

本文所采用的基于路径的预测方法包括 4 种: SP 算法; Katz[11]算法; RWR[13]算法; FriendLink[12]算法。

3.3　科研合作推荐模型的评价方法

采用三种评价指标验证算法的推荐性能，即平均正确率均值(Mean Average Precision, MAP)[18]和曲线下的面积(Area Under Curve, AUC)[19]。在合作者推荐的情景中，对于潜在的科研合作对象，根据推荐模型的推荐结果与现实情况进行比较，最终可能会出现 4 种结果，如表 2 所示。

表 2　推荐模型产生的结果矩阵

	推荐	不推荐
有合作	TP	FN
无合作	FP	TN

准确率的计算公式如公式(4)所示。

$$Precision = \frac{|TP|}{|TP| + |FP|} \quad (4)$$

MAP 的计算公式如公式(5)所示。

$$MAP = \frac{1}{n}\sum_{u=1}^{n}\sum_{k=1}^{r_u} Precision_u@k \quad (5)$$

其中，u 是目标节点，r_u 表示与目标节点 u 相关的节点，$Precision_u@k$ 表示当向目标节点 u 推荐 top-k 个节点时的准确率。MAP 不仅可以说明推荐算法的准确率还能揭示推荐算法的排序能力，因此使用 MAP 可以更好地对推荐算法的整体性能进行评估。

此外，本文使用曲线下面积(AUC)作为评价指标。假设两个节点之间当前没有链接，但是将来会产生链接，这类链接被称为“缺失链接”；假设两个节点现在

不存在链接，将来也不会产生链接，这类链接被称为“错误链接”。AUC 实际上就是比较缺失链接的得分和错误链接的得分。在每次实验中分别选择一条缺失链接和一条错误链接进行分数比较，该实验独立进行 n 次，假设其中有 n' 次缺失链接的得分大于错误链接的分数，有 n''次缺失链接的得分和错误链接的得分相等，则 AUC 值可以由公式(6)获得。AUC 的取值越大，说明推荐效果越好。

$$AUC=\frac{n'+0.5\times n''}{n} \tag{6}$$

3.4 个人科研合作推荐的实验结果与讨论

(1) 基于网络结构特征的个人科研合作推荐结果

在个人层次的科研合作推荐上，本文使用 8 种基于网络结构特征的链接预测算法，即基于邻居节点的算法(AA、CN、Jaccard 和 PA)和基于路径的算法(FL、RWR、Katz 和 Path_Sim)。在 Katz 算法中，设定 β 为 0.001。鉴于本文所构建的科研合作者网络最大直径为 18，为了计算的可行性，设定 k 等于 10，即 k>10 的情况不予考虑。在 FriendLink 算法[12]中，根据“六度分割”理论设定的 l 值在[2,5]，该论文最终的实验结果表明 l 取 3 时，算法能够获得最好的预测效果。鉴于此，本文设置 l 为 3。

图 1 显示了 4 种基于邻居节点的算法结果，可以看出，AA 算法在 MAP 和 AUC 上都获得了最高值；CN 算法和 PA 算法在 AUC 上的表现较为接近，CN 算法的 AUC 值明显高于 Jaccard 算法；PA 算法的 MAP 值很低，AUC 值相对要高一些。图 2 显示了 4 种基于路径的算法在各项评价指标上的表现结果，可以看出，在 MAP 值上 RWR 算法得分最高，其次是 FriendLink 算法,Path_Sim 则得分最低；在 AUC 上 RWR 和 Path_Sim 的效果优于其他两个算法。

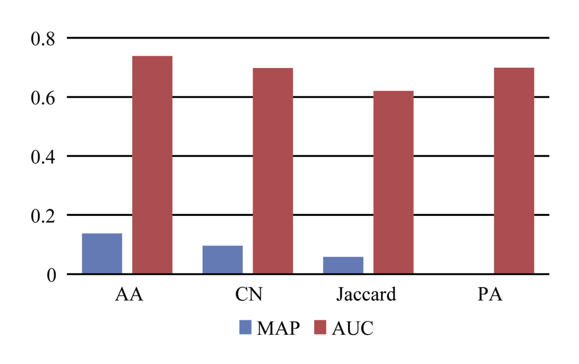


图 1　基于邻居节点的预测算法结果——作者层次

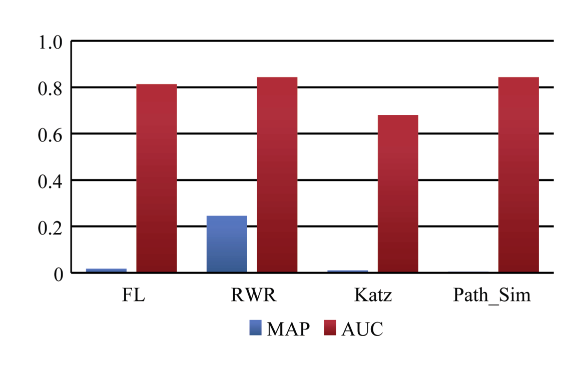


图 2　基于路径特征的预测算法结果——作者层次

(2) 基于特征融合的个人科研合作推荐结果

综合各种算法在三个评价指标上的表现，最终选择 AA 作为基于邻居节点的算法代表，选择 RWR 作为基于路径的算法代表。

对融合参数 α 进行调整的折线图如图 3 所示。

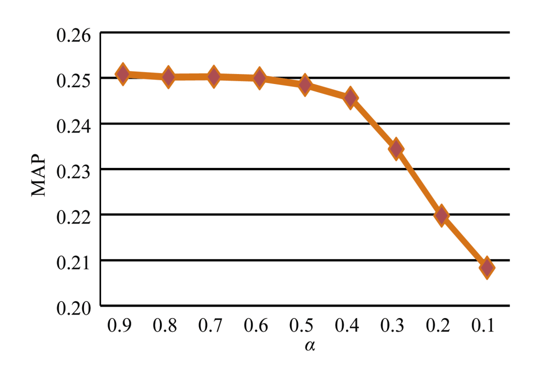


图 3　融合参数 α 调整折线图——作者层次

由图 3 可以看出，当 α 取值为 0.9 时，融合模型的结果最好。因此取 α 为 0.9，利用公式(3)将这两项特征融合，利用融合模型(PHM)得到最终的推荐结果，如图 4 所示。可以看出，无论是 MAP 还是 AUC 值，PHM 模型均获得最佳性能，这验证了将多种特征融合在一起进行科研合作推荐能够获得更好效果的设想。

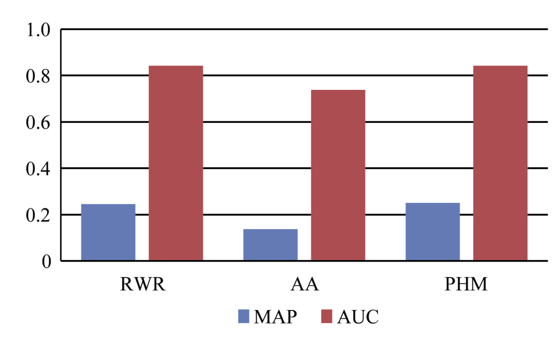


图 4　各种链接预测算法的结果比较——作者层次

笔者利用从金融领域采集的 2000 年–2014 年的论文，采用 PHM 模型推荐可能的合作者。表 3 显示了推荐的部分科研合作者。通过对科研作者的机构进行比对分析，发现许多作者对处于相同或地理上邻近的机构。例如，在第一对推荐结果中，阎庆民曾任职于中国银行业监督管理委员会(银监会)，谢平则受雇于中国投资公司(在地理位置上与银监会较为接近)。在第 7 对推荐结果中，胡浩和樊志刚都任职于中国工商银行。研究结果也进一步验证了 Evans 等关于机构和地理因素对科研合作的影响的论证[20]。后者的研究表明，受制于机构和地理位置等限制因素，科学家更倾向于建立内部机构的合作；对于作者机构以外的选择，更愿意寻求那些在地理上更为接近的机构来进行合作。模型推荐结果从机器学习的角度为科研合作往往是建立在同一个或地理上密切的机构之间的作者之上这一论点提供了新的依据。

表 3 PHM 模型推荐的部分科研合作作者

编号	作者 1	作者 2
1	阎庆民	谢 平
2	陈卫东	姜波克
3	姜波克	张健华
4	阎庆民	陈卫东
5	阎庆民	张健华
6	温信祥	樊志刚
7	胡 浩	樊志刚
8	王佳佳	樊志刚
9	张燕生	唐 旭
10	胡 浩	马素红

3.5 机构科研合作推荐的结果与讨论

在机构层次使用 8 种基于网络结构特征的链接预测算法，即基于邻居节点的算法(AA、CN、Jaccard 和 PA)和基于路径的算法(FL、RWR、Katz 和 SP)作为基线方法。各算法的参数设置与个人层次的参数设置相同。图 5 显示了基于邻居节点的链接预测算法对机构合作关系进行预测的结果，可以看出，综合考虑 MAP 和 AUC 值，CN 算法在 4 种算法之中表现最好。如图 6 所示，在基于路径的算法中 RWR 算法的 MAP 值最高，综合表现最优；SP 算法的 MAP 值也比较高，但是 AUC 值低于 RWR 算法；Katz 算法和 FriendLink 算法的两项评价指标均较弱。

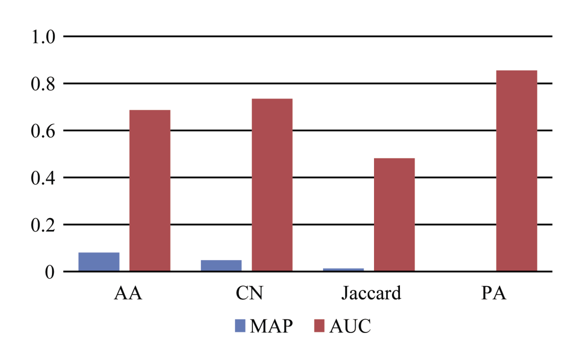


图 5 基于邻居节点的链接预测算法结果比较——机构层次

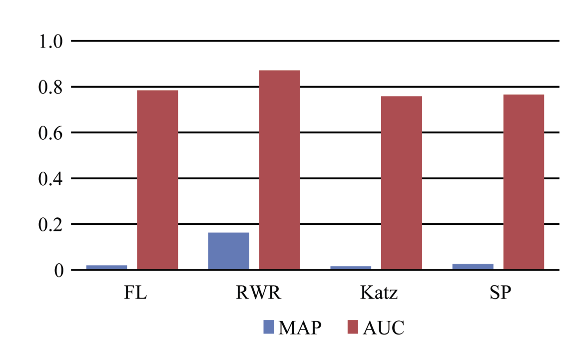


图 6 基于路径特征的预测算法结果比较——机构层次

根据上述各种算法的 MAP 指标比较，笔者选择 RWR 和 CN 两种算法进行融合，进行合作机构推荐。对区域层次 α 参数的调整如图 7 所示，选定 α 为 0.7。图 8 显示了融合算法与其他算法进行比较的结果。可以看出，在机构层面，融合算法能够获得更好的推荐结果。

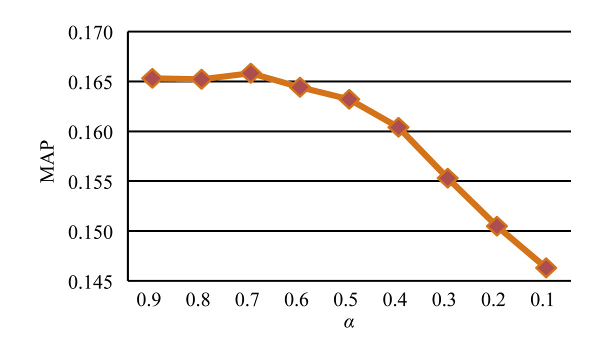


图 7 融合参数 α 调整折线图

利用从金融领域采集的 2000 年–2014 年的论文，采用 PHM 模型推荐可能的合作机构，部分预测的科研合作机构如表 4 所示。

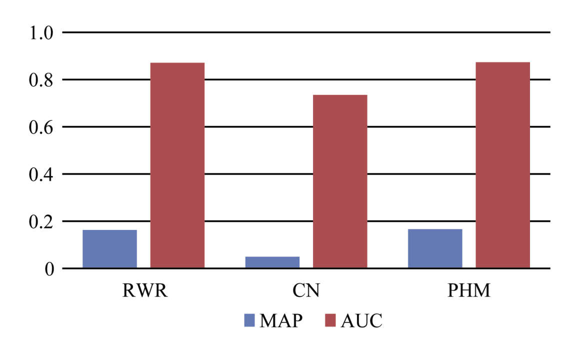


图 8　链接预测算法比较——机构层次

从表 4 可以看出，PHM 模型的推荐结果在一定程度上反映了科研机构在地理位置上的近邻性。例如，第 3、5、6 和 8 组的科研合作机构是隶属于同一所大学的不同学院；第 7 和 9 组则位于同一省份或者城市。除了地理因素，推荐结果还反映了研究机构之间的主题相似性。通过对研究机构的主题进行分析发现，第 1、4 和 10 组机构具有很高的主题相似性。

表 4　PHM 模型推荐的部分科研合作机构

编号	机构 1	机构 2
1	中国金融学会金融史专业委员会	上海市金融学会
2	烟台大学经管学院	东北财经大学公共管理学院
3	云南财经大学商学院	云南财经大学会计学院
4	云南大学国际关系研究院	南开大学日本研究院
5	复旦大学管理学院产业经济系	复旦大学管理学院财务金融系
6	东北财经大学应用金融学院	东北财经大学职业技术学院
7	西南大学地理科学学院	重庆大学建设管理与房地产学院
8	华东师范大学俄罗斯研究中心	华东师范大学国际关系与地区发展研究院
9	浙江大学理学院	浙江水利水电专科学校
10	中国科学技术大学公共事务学院	西南科技大学政治学院

3.6　区域科研合作推荐的结果与讨论

在区域层次，采用与机构合作层次相同的 8 种链接预测算法。在 Katz 算法中，设定 β 为 0.001，统计发现区域合作网络的直径为 3，平均路径长度为 1.5，为了计算的可行性，设定 k 等于 3；在 FriendLink 算法中，设定 l 等于 2。

在区域层次上也使用相同的 8 种基于网络结构特征的链接预测算法，即基于邻居节点的算法(AA、CN、Jaccard 和 PA)和基于路径的算法(FL、RWR、Katz 和 SP)作为基线方法。

图 9 显示了基于邻居节点的链接预测算法对于机构合作关系预测的结果，可以看出，综合考虑 MAP 和 AUC 值，CN 算法在 4 种算法之中表现最好，Jaccard 算法和 PA 算法 MAP 值和 AUC 值相当，AA 算法表现相对不好。如图 10 所示，在基于路径的算法中 RWR 算法的 MAP 值最高，在两个评估指标上综合表现最优；SP 算法的 MAP 值也比较高，但是 AUC 值低于 RWR 算法；Katz 算法和 FriendLink 算法的两项评价指标相对较弱。

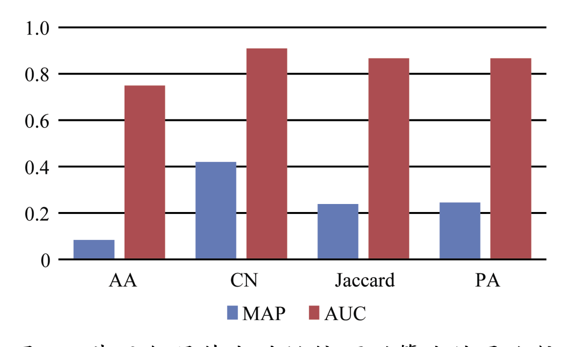


图 9　基于邻居节点的链接预测算法结果比较——区域层次

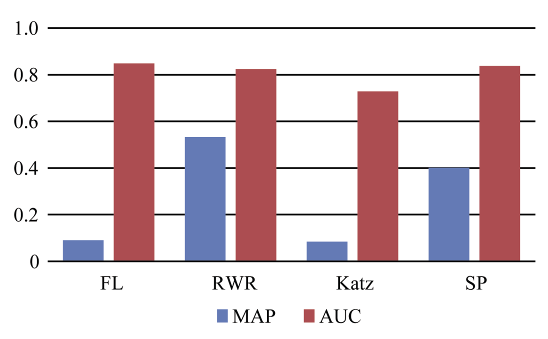


图 10　基于路径特征的预测算法结果比较——区域层次

根据上述各种算法的 MAP 值比较，选择 RWR 和 CN 两种算法进行融合，进行合作区域推荐。融合参数 α 的调整如图 11 所示，最终选定 α 为 0.5。图 12 显示了融合算法与其他算法进行比较的结果。可以看出，在区域层面，融合算法能够获得更好的推荐结果。

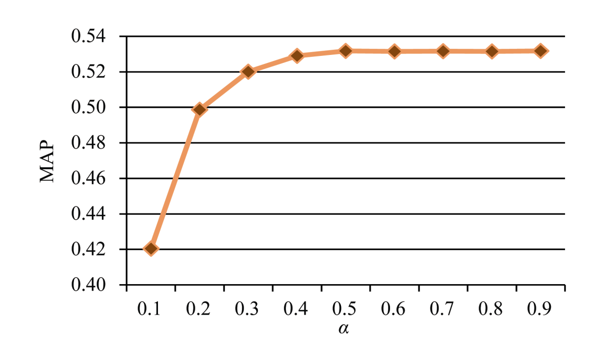


图 11 融合参数 α 调整折线图

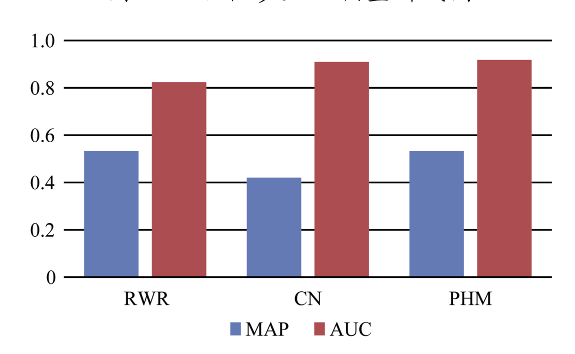


图 12 链接预测算法比较——区域层次

通过使用提出的融合方法，笔者尝试利用 2000 年–2014 年的区域合作数据来预测的潜在的区域合作对，其结果如表 5 所示。

表 5 PHM 模型推荐的部分科研合作区域

编号	区域 1	区域 2
1	陕西	海南
2	河北	重庆
3	重庆	陕西
4	黑龙江	重庆
5	天津	广西
6	吉林	重庆
7	四川	辽宁
8	江苏	广西
9	贵州	天津
10	海南	江苏

从表 5 可以看出，在区域层面，模型的推荐结果对物理位置的近邻性的反映程度弱于作者和机构层面。通过对这些地区的 GDP 进行深入分析发现，模型的推荐结果更多地反映了区域之间经济状况的差异性。例如，在第 1 对推荐结果中，陕西的 GDP 为 1 兆 7 689 亿 9 400 万元人民币(2014 年)，而海南则仅为 3 500 亿 7 200 万元人民币(2014 年)。在第 2 对推荐结果中，河北的 GDP 为 2 兆 9 421 亿 4 000 万元人民币，而重庆则仅为 1 兆 4 265 亿 4 000 万元人民币。在区域层面，模型更倾向于推荐经济发展互补的区域进行合作。

3.7 个人、机构与区域科研合作推荐的综合分析

对比 4 种基于邻居节点的算法(AA、CN、Jaccard 和 PA)和 5 种基于路径的算法(FL、RWR、Katz、SP 和 Path_Sim)在个人、机构以及区域层次的科研合作推荐结果，发现不同算法在各层次的科研合作网络中的推荐效果呈现多样性。例如，在基于路径的预测方法中，RWR 算法在个人、机构和区域三个层次上的链接预测中均展现出最好的结果；在基于邻居节点的链接预测算法中，AA 算法在区域层次表现最好，CN 算法在区域和机构层次表现得最好。对比基于网络结构的方法(AA、CN、Jaccard、PA、FL、RWR、Katz、SP 和 Path_Sim)、融合方法(PHM)，发现无论是 MAP 还是 AUC 值，融合模型均获得最佳性能，这表明将多种特征融合在一起能够有效提升推荐效果。

本文以金融领域的科研合作网络作为实证研究对象，将科研合作对象(包括个人、机构和区域等)抽象成为一个节点，将科研对象之间的合作关系抽象成为一条边。由于网络拓扑结构在不同学科门类下的分布呈现较高的一致性[21]，与学科门类并无实质关联，所以本文提出的模型和方法理论上可以应用到其他学科。值得说明的是，个人、机构和区域三个层次的科研合作推荐研究并非各自孤立，而是彼此关联。研究个人层面的科研合作推荐有助于发现科研人员合作的现状，揭示影响合作的微观因素，例如研究主题的差异性、研究机构的同一性等；研究机构层次的科研合作推荐有助于发现科研团队合作的现状，揭示影响合作的中观因素，例如地理位置的远近等；研究区域层次的科研合作推荐则可以更多地发现影响合作产生的宏观因素，例如区域经济发展的差异性等。从推荐结果来看，在个人层面，算法倾向于推荐具有相同机构的作者进行合作；在机构层面，倾向于推荐地理位置较

为靠近的机构进行合作；在区域层面，则倾向于推荐经济发展互补的区域进行合作。尽管算法推荐的结果并不能作为实际工作的指导，但上述工作对于科研人员、机构与区域选择恰当的合作对象、促进学术交流与提高科研产出仍然提供了新的思路，具有一定的启发与借鉴意义。

4 结 语

本文通过利用影响科研合作关系产生的因素——网络结构特征(包括邻居节点和路径)，构建了多种特征融合的科研合作推荐方法，并对金融领域的个人、机构与区域层次的科研合作推荐进行了实证研究。在网络结构特征方面，检验了 4 种基于邻居节点的算法(AA、CN、Jaccard 和 PA)和 5 种基于路径的算法(FL、RWR、Katz、SP 和 Path_Sim)对于个人、机构和区域层次的科研合作推荐的效果，发现不同算法适用于不同的网络，在各网络中的推荐效果呈现多样性。本文提出的将邻居节点特征、路径特征进行融合的推荐模型在 MAP 与 AUC 指标上都取得了比只考虑其中一项特征更好的效果，这表明融合多种特征的推荐算法优于仅仅考虑单一特征的推荐算法。对于机构和区域层次的科研合作推荐则发现对于机构合作来说，影响合作产生的主要因素是地理位置，而影响区域合作的主要因素是经济发展水平。其研究发现有助于科研人员、机构与区域选择恰当的合作对象，促进学术交流与合作。后续还将考虑更多的特征来丰富各层次实体的科研合作推荐理论与方法。

参考文献：

[1] Dai C, Chen L, Li B, et al. Link Prediction in Multi-relational Networks Based on Relational Similarity[J]. Information Sciences, 2017, 394-395: 198-216.

[2] Moradabadi B, Meybodi M R. Link Prediction Based on Temporal Similarity Metrics Using Continuous Action Set Learning Automata[J]. Physica A: Statistical Mechanics & Its Applications, 2016, 460: 361-373.

[3] Yu C, Zhao X, An L, et al. Similarity-based Link Prediction in Social Networks: A Path and Node Combined Approach[J]. Journal of Information Science. DOI: 10.1177/0165551516664039.

[4] 张斌，马费成．科学知识网络中的链路预测研究述评[J]．中国图书馆学报，2015，41(3)：30-47．(Zhang Bin, Ma Feicheng. A Review on Link Prediction of Scientific Knowledge Network[J]. Journal of Library Science in China, 2015, 41(3): 30-47.)

[5] Lu L, Zhou T. Link Prediction in Complex Networks: A Survey [J]. Physica A: Statistical Mechanics and Its Applications, 2011, 390(6): 1150-1170.

[6] Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and Accurate Link Prediction in Social Networking Systems [J]. Journal of Systems and Software, 2012, 8(5): 2119-2132.

[7] Chen J, Geyer W, Dugan C, et al. Make New Friends, but Keep the Old: Recommending People on Social Networking Sites [C]//Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI). 2009: 201-210.

[8] Adamic L,Adar E. How to Search a Social Network [J]. Social Networks, 2005, 27(3): 187-203.

[9] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining [M]. The 1st Edition. Boston: Addison Wesley, 2005: 65-84.

[10] Costa L da F, Rodrigues F A, Travieso G, et al. Characterization of Complex Networks: A Survey of Measurements[J]. Advances in Physics, 2007, 56(1): 167-242.

[11] Katz L. A New Status Index Derived from Scientometric Analysis[J]. Psychometrika, 1953, 18(1): 39-43.

[12] Papadimitriou A, Symeonidis P, Manolopoulos Y. Fast and Accurate Link Prediction in Social Networking Systems [J]. Journal of Systems and Software, 2012, 8(5): 2119-2132.

[13] Pan J, Yang H, Faloutsos C,et al. Automatic Multimedia Cross-modal Correlation Discovery [C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Ming, Seattle, WA,USA. 2004: 653-658.

[14] Yan E, Guns R. Predicting and Recommending Collaborations: An Author-, Institution-, and Coutry-level Analysis [J]. Journal of Informetrics, 2014(8): 295-309.

[15] 张斌．科研合作网络中的链路预测研究[D]．武汉：武汉大学，2016．(Zhang Bin. Research on Link Prediction of Scientific Collaboration Network [D]. Wuhan: Wuhan University, 2016.)

[16] 刘萍，郑凯伦，邹德安．基于 LDA 模型的科研合作推荐研究[J]．情报理论与实践，2015，38(9)：79-85．(Liu Ping, Zheng Kailun, Zou De'an. Research on Cooperative Recommendation Based on LDA Model[I]. Information Studies: Theory & Application, 2015, 38(9): 79-85.)

[17] 吕伟民，王小梅，韩涛．结合链路预测和 ET 机器学习的科研合作推荐方法研究[J]．数据分析与知识发现，2017，1(4)：38-45．(Lv Weiming, Wang Xiaomei, Han Tao. Recommending Scientific Research Collaborators with Link Prediction and Extremely Randomized Trees Algorithm[J].

Data Analysis and Knowledge Discovery, 2017, 1(4): 38-45.)

[18] Yue Y, Finley T, Radlinski F, et al. A Support Vector Method for Optimizing Average Precision[C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, Netherlands. New York: ACM, 2006:271-278.

[19] Zhou T, Lv L, Zhang Y C. Predicting Missing Links via Local Information [J]. European Physical Journal B - Condensed Matter and Complex Systems, 2009, 71(4): 623-630.

[20] Evans T S, Lambiotte K, Panzarasa P. Community Structure and Patterns of Scientific Collaboration in Business and Management[J].Scientometrics, 2011, 9(1): 381-396.

[21] 巴志超，李纲，朱世伟. 基于知识超网络的科研合作行为实证研究和建模[J]. 情报学报，2016，35(6): 630-639. (Ba Zhichao, Li Gang, Zhu Shiwei. Empirical Study and Modeling of Scientific Cooperation Behavior Based on Knowledge Hypernetwork[J]. Journal of the China Society for Scientific and Technical Information, 2016, 35(6): 630-639.)

作者贡献声明:

余传明：设计并实施技术方案、技术路线，提出论文主要研究思路，优化研究方案及技术路线，论文修改；

龚雨田，赵晓莉：实验操作，论文初稿撰写；

安璐：提出研究思路，数据分析，论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yuchuanming2003@126.com。

[1] 余传明，龚雨田，赵晓莉，安璐. data.zip. 个人、机构、区域三个层面的科研合作数据.

收稿日期: 2017-05-31

收修改稿日期: 2017-07-18

Collaboration Recommendation of Finance Research Based on Multi-feature Fusion

Yu Chuanming[1] Gong Yutian[1] Zhao Xiaoli[1] An Lu[2]

[1](School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China)

[2](School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] Research collaboration builds an important social network system. This paper proposes a new recommendation model for research collaboration in finance, aiming to promote the scientific collaboration and improve research productivity. [**Methods**] First, we established the scientific collaboration networks at individuals, institutions and regions levels. Then, we established a recommendation model based on network neighbors and paths. Finally, we conducted empirical study to examine the model at three levels. [**Results**] A total of 68 905 articles published from 2000 to 2014 on finance were analyzed to construct their research collaboration networks. The AUC values of the proposed model at individual, institutional and regional levels were 84.25%, 87.34%, and 91.84%, respectively, which were higher than those of the traditional algorithms. [**Limitations**] The training and testing sets were only classified by time. More segmentation methods were needed to optimize the new model. [**Conclusions**] This study helps researchers find collaboration opportunities, and provides new directions for studies on scientific collaboration networks.

Keywords: Link Prediction Scientific Collaboration Recommendation Scientific Collaboration Network Multi-feature Fusion